Ecological Informatics 38 (2017) 76–81

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

The use of jackknifing for the evaluation of geographic profiling reliability

Alessio Papini [a,*], D. Kim Rossmo [b], Stephen C. Le Comber [c], Robert Verity [c], Mark D. Stevenson [c], Ugo Santosuosso [d]

[a] Department of Biology, Università di Firenze, Via Micheli, 3, 50121 Firenze, Italy
[b] Center for Geospatial Intelligence and Investigation, School of Criminal Justice, Texas State University, San Marcos, TX, USA
[c] Department of Organismal Biology, School of Biological and Chemical Sciences, Queen Mary University of London, London, UK
[d] Department of Clinical and experimental Medicine, Università di Firenze, Largo Brambilla, 3, 50134 Firenze, Italy

ARTICLE INFO

Article history:
Received 14 October 2016
Received in revised form 29 January 2017
Accepted 2 February 2017
Available online 4 February 2017

Keywords:
Geographic profiling
Jackknife
Bootstrap
Biological invasions
Caulerpa

ABSTRACT

The use of geographic profiling (GP), based on "Rossmo's formula", a technique derived from criminology, has been previously proven to be effective in assessing the origin of invading species. The application on *Caulerpa taxifolia* showed the most probable center of spread of the invasion. This article discusses a method of assessing the degree of robustness of the results obtained with Rossmo's method.

To provide an evaluation of the reliability of geographic profiling results we used the jackknife technique, randomly eliminating part of the data set for a given number of replicates (500) in order to analyze the obtained result for each replicate. In GP the results are a series of images with geoprofiling prioritization, each produced with one of the replicates. These images can be summarized in three different ways: (1) OR, depicting all the high probability pixels from the series of replicates; (2) AND, depicting only those high probability pixels present in every replicate; and (3) MEAN, depicting the mean color value for each pixel calculated from all the replicates. We show that jackknifing can be a useful method to increase robustness of GP analysis in criminology, epidemiology and biological invasions. Summarizing jackknifing results with the OR logical operator yields the highest sensitivity and worst specificity, while the use of the AND operator increases specificity but reduces sensitivity. Using the mean of the pixel values maintains the visualization of the areas of highest priority (specificity), while also showing the surrounding area with varying colors, analogous to confidence limits.

1. Introduction

1.1. Generality about geographic profiling

Geographic profiling (GP) is an analytical technique used in criminology, with the aim of calculating the most probable origin of linked crimes, which is usually the offender's home. GP is used by police forces around the world to help focus investigations and prioritize suspects in cases of serial crimes (Rossmo, 2012).

GP input consists of spatial data about the locations of linked crimes, which is used to create a probability surface to overlay on the map of interest in the form of a geoprofile (Rossmo, 2000). GP does not provide an exact origin location, but, instead, provides a prioritization pattern for investigations based on a descending order of the probability height on the geoprofile (Rossmo, 2000).

The model is based on two components: a distance-decay function, such that the probability of a crime (or other events with a localization on a map) decreases with increasing distance from the offender's residence; and a buffer zone, within which the probability increases with distance (Rossmo, 2000). The distance-decay function is due to travel costs – both for human criminals and invasive species – (Stevenson et al., 2012), in economical or energy terms, respectively. The buffer zone is linked to the avoidance by criminals of locations too close to their residence. In biology, the existence and the extent of the buffer zone should be analyzed case by case. For example, Dramstad (1996), Saville et al. (1997), Singh et al. (2001) and Stevenson et al. (2012) showed evidence of a buffer zone in trees and bees.

1.2. Geographic profiling applied to biological invasions

GP has been used in biology to analyze the origins of infectious diseases (Le Comber et al., 2011; Verity et al., 2014), to predict the locations of multiple nest locations of bumble bees (Suzuki-Ohno et al., 2010), and to study the patterns of animal foraging (Le Comber et al., 2006; Raine et al., 2009) and great white sharks predating on seals (Martin et al., 2009).

Stevenson et al. (2012) used GP to identify the origin of the invasion of a species, starting from the current known locations of their populations. The places colonized by the invasive populations were considered

[*] Corresponding author.
E-mail address: alpapini@unifi.it (A. Papini).

http://dx.doi.org/10.1016/j.ecoinf.2017.02.001
1574-9541/© 2017 Elsevier B.V. All rights reserved.

analogous to crime sites, while the source or sources of the invasion were considered analogous to the criminal's home. The same authors tested various parameters of GP, also in comparison to other spatial techniques, such as the center of minimum distance, the spatial mean, the spatial median and a single parameter density model. GP gave better results compared to the other techniques in 52 of the 53 data sets explored for invasive species in Great Britain. Stevenson et al. (2012) provided a list of values for the buffer zone radius evaluated as the most appropriate in their analysis. The technique was applied with success on biological invasions of algae (Papini et al., 2013) and insects (Cini et al., 2014).

Invasive species are considered to be one of the main causes of biodiversity loss (Vitousek et al., 1996; Wilcover et al., 1998). They can damage native species through predation and competition, by modifying ecosystem functions by altering the abiotic environment and by spreading pathogens (Strayer et al., 2006; Pimentel et al., 2005).

The invasion of macroalgae, such as some species belonging to genus *Caulerpa* (Caulerpales, Chlorophyta), is one of the main threats to marine natural environments (Meinesz et al., 2001). Two species of *Caulerpa* J. V. Lamouroux, *Caulerpa taxifolia* (Vahl) C. Agardh and *C. racemosa* (Forskål) J. Agardh var. *cylindracea* (Sonder, 1845) Verlaque, Huisman and Boudouresque, caused severe biological pollution (Piazzi et al., 2005; Verlaque et al., 2003, 2004) in the Mediterranean. One interesting feature of the invasion by the first species, is that the origin is known - an accidental release from the aquarium of Monaco in 1984 (Meinesz and Hesse, 1991; Meinesz et al., 2001). Geographic profiling of the invasive caulerpas spread in the Mediterranean was already used with success by Papini et al. (2013), taking advantage of the fact that the spreading origin of *Caulerpa taxifolia* is known, and the related data set is a good starting point for calibrating the technique.

The GP analysis applied to biological invasions has certain limitations: the results obtained with Rossmo's formula are based on the postulate that all spreading events should come from one or a limited number of origins, which is not always true for biological invasions where secondary sites of invasion may frequently derive from the original primary or more independent introductions may occur (Santosuosso and Papini, 2016). Furthermore, vegetative propagation and other “slow” ways of environmental spread may obscure the general pattern (Papini et al., 2013; Verity et al., 2014).

An alternative approach may be a series of geoprofiles using different time periods for the data (e.g., year 1, year 2, year 3, etc.), a technique already used with success in crime analysis (Rossmo and Velarde, 2008). Such an approach allows for a better understanding of the spread from secondary sites, reducing the “noise” in the data. This was the approach taken by Stevenson et al. (2012), who fitted the parameters of the model using a maximum likelihood approach from a time series. Moreover, almost certainly some of the sites of the invading algae are unknown, making the final result approximate (Papini et al., 2013). This is also frequently true in criminology where the accuracy of data used for GP may affect the accuracy of the analysis (Rossmo, 2005; Snook et al., 2005).

1.3. Jackknifing and bootstrapping techniques

A possible method to analyze the effect of the errors derived from the limitations linked to the geoprofile may be the use of data resampling techniques such as jackknifing or bootstrapping (Miller, 1974; Efron, 1979, 1982; Efron and Tibshirani, 1986, 1993). Both methods are commonly used in other biological analyses (Manly, 2006); for example, bootstrapping is commonly used to assess the robustness of phylogenetic analysis (Felsenstein, 1985). The two methods are very similar, consisting both in a random deletion of part of the data, with the jackknife using such a reduced data set, while the bootstrap substitutes the deleted data by duplicating some of the remaining data items (Meyer et al., 1986).

The aim of this study, was to use the jackknife technique (Efron, 1979; Miller, 1974) to test the robustness of a GP analysis and to provide a framework on how to summarize the resulting graphical results. After van Belle et al. (2004), a statistical or an analytical procedure is robust if it performs well when the needed assumptions are not violated “too badly". After the same authors, it is not a strictly mathematical definition, but robustness should provide a measure of the confidence limits of the obtained results. Even Umeton et al. (2011) defined robustness as “Robustness is the persistence of a system property respect to perturbations”. In the case of geographic profiling, the jackknife technique should provide an idea of the robustness of the analysis and of the confidence within which we can look for the point of origin of the sites of biological invasion on a map. The assessment of the confidence limits of geographic profiling may be extended to other analyses, including those outside the field of biological invasions.

For assessing robustness, it is possible to create new data sets simply by resampling the observed data. Such an analysis requires taking a series of subsamples from a data set a given number of times. Some observations appear once, others twice, others not at all (van Belle et al., 2004). In jackknifing, a part of the sample is systematically omitted, for example, by removing one data point at a time, and the analysis is then carried out for each newly constructed subset (Efron, 1982; Efron and Tibshirani, 1993).

2. Materials and methods

2.1. Theory

The model for geoprofiling analysis was described by Rossmo (2000), who compared the use of Manhattan and Euclidean distances, preferring the former to describe criminal movement in urban areas. However, Le Comber et al. (2006) and Stevenson et al. (2012) suggested that Euclidean distances are more appropriate for animal and plant movements in nature.

The geographic profiling function generates a surface where each pixel has a different priority score indicating the optimal search pattern for the sources of invasive species (Stevenson et al., 2012). For each pixel with coordinates (i, j) of the target area, the score function (p) is calculated as follows (Rossmo, 2000):

$$p_{ij} = k \sum_{n=1}^{C} [\phi_{\iota\varphi} / (| x_i - x_n | + | y_j - y_n |)^{f} + (1 - \phi)(B^{g-f})/(2B - | x_i - x_n | - | y_j - y_n |)^{g}]$$

where $\phi_{\iota\varphi}$ is equal to 1 if $|x_i - x_n| + |y_j - y_n| > \mathrm{B}$; otherwise, it is equal to 0.

This formula representation uses the Manatthan metric: “$|x_i - x_n| + |y_j - y_n|$”.

For point p with coordinates (i, j), the formula sums the probability across all the locations where the invading organism was found. After Rossmo (2000), Φ functions as a switch that is set to 0 for sites within the buffer zone, and 1 for sites outside the buffer zone. k is an empirically determined constant, which was set to 1 in our study. B is the radius of the buffer zone, and C is the number of events (in this case the reports about the presence in a given locality of the invader). f and g are parameters that control the shape of the distance-decay function on either side of the buffer zone radius. For our analysis we used the same parameters as in Papini et al. (2013), calibrated on the known origin of *C. taxifolia*. The parameters specify the increase in dispersal probability moving away from the source, reaching a maximum value at a distance equal to the radius of the buffer zone. This reflects the reduced probability of dispersal within the buffer zone and the fact that dispersal probability declines with distance (Stevenson et al., 2012).

This function produces a search priority surface for the inputted locations on the user-provided map (Rossmo, 2000). Rossmo described the equation as a curve, which, when plotted in three dimensions, resembles the shape of a volcano with a caldera. The sum of these